Socially Relevant Identity: Addressing Selection Bias Issues and Introducing the AMAR (All Minorities at Risk) Data.[1]

Jóhanna K Birnir
Univeristy of Maryland
David D Laitin
Stanford University
Jonathan Wilkenfeld
University of Maryland
Agatha Hultquist
University of Maryland
David M. Waguespack
University of Maryland
Ted R. Gurr
University of Maryland

Abstract: The paper introduces the AMAR (All Minorities at Risk) data, a coded sample of socially recognized and salient ethnic groups. We describe the data and review under-explored selection issues arising with truncated ethnic group data, especially when moving between levels of analysis. Next we suggest some directions for the future study of ethnicity and conflict using our bias corrected data, including a better estimate of overall group propensity for ethnic violence. We also correlate group violence and some prominent group and country level variables proposed as causes of ethnic violence. Our correlations suggest that some group level relationships likely are missed and/or incorrectly specified in the literature. Furthermore, country level measures such as ethnic heterogeneity and economic development, while related to *absolute levels of violence* in a given country, in and of themselves may not be as significant correlates of *ethnic group propensity for rebellion* as has been previously reported.

[1] We thank the National Science Foundation for supporting this work (grant #SES0718957). Parts of the work have been presented at Juan March Institute, Folke Bernadotte Academy, Penn State University, Uppsala University, International Studies Association, University of California Los Angeles, Stanford University, University of Maryland, Midwest Political Science Association, the Pentagon and Yale University. We thank the many discussants and others who have generously read and commented including James Fearon, Andreas Wimmer, Stephen Saideman, Lars-Erik Cederman, and Kathleen Cunningham.

Introduction

This paper addresses the well-known selection bias issue plaguing the Minorities at Risk (MAR) dataset that has nonetheless been widely used to examine the association between ethnic diversity and violent ethnic political mobilization. Precise measurement of this association has been challenging[2] due in large part to the absence of a group-level *sample* - free from known selection issues - that would allow us to estimate the probability of any ethnic group to be engaged in violent confrontation with the state.

The paper introduces the AMAR (All Minorities at Risk) *sample* of socially recognized and salient ethnic groups, which we call the AMAR Phase I data. Guided by theories of ethnic politics that help drive the selection of the appropriate sampling frame, the AMAR sample frame[3] (Birnir et. al 2015) enumerates 1202 ethnic groups, including over 900 groups that were not included in the MAR groups data project. [4] From this set of new groups, we code in this paper a random sample of 74 groups, stratified by region and size, for the suite of extant MAR variables. With statistical weighting, we combine this random set with the current MAR data, allowing us to

[2] For an overview of the literature on the detrimental effects of ethnicity see Chandra (2012) chapter 1.

[3] For an extended discussion of the challenges to constructing a sample frame for ethnic see Birnir et al. (2015)

[4] The total number of groups in the current AMAR sample frame differs from the total number listed in Birnir et al. 2015, which was 1196, because six new groups were added since the paper's publication. These new groups were added based on updated information, and include the Afromexicans in Mexico, Bantenese in Indonesia, Bemba/Shila in the Democratic Republic of Congo, French in Belgium, Italians in Germany, and Irish in the United Kingdom.

address selection bias concerns[5] that have been a nemesis for existing studies of the relationship between ethnicity and violence.

In this paper, we first review the selection concerns in the study of ethnic conflict with an emphasis on the underexplored selection issue that arises with truncated data especially when moving between levels of analysis. Next we describe our sampling solution and the resulting AMAR data (taking into account the impact of error in the sampling frame) with an eye to assessing the prevalence of group participation in violence. Finally, to illustrate issues of selection bias and suggest some directions for the future study of ethnicity using our bias corrected data, we use the coded sample to better estimate overall group propensity for ethnic violence in the world. With that goal, we correlate group violence and some prominent group level and country level variables that have been proposed as causes of ethnic violence, including political, economic, and cultural grievances, group concentration, wealth, and ethnolinguistic fractionalization.

Substantively, along the lines of Fearon and Laitin (1996), but with results for the entire world, our descriptive findings suggest that only a minority of widely recognized ethnic groups ever engage in conflict against the state. The preliminary group level correlations support the concerns in the literature that selection bias decreases the likelihood that relationships are detected. Moreover, our suggestive correlations using data collected at different levels indicate that ethnic heterogeneity and economic development, while related to *absolute levels of violence* in a given country, in and of themselves may not be as significant correlates of *ethnic group propensity for rebellion* as has been previously reported (Reagan and Norton 2005; Olzak 2006; Walter 2006; Cetinyan 2002). In sum, the AMAR sample data permits estimations in future

[5] For a discussion of the selection bias in MAR data see Fearon & Laitin, 1996; Fearon & Laitin, 2002, 2003; Fearon, 2003; Öberg, 2002a; Hug, 2003, 2013; Birnir, 2007; Brancati, 2006, 2009.

research of conflict potential at the group level with a greater degree of confidence in our results.

In order to make valid inferences about ethnic conflict, there is an urgency in addressing the group level data used in the study of ethnic conflict. As social scientists in many fields have sought to understand the mechanisms underlying their causal claims, they have found cross-country regressions to be unhelpful, and have sought greater levels of disaggregation that would permit better controls and easier identification. While progress was fruitful in earlier studies of ethnic rebellion relying on the country/year as the unit of analysis (Collier and Hoeffler 2004; Fearon and Laitin 2003), critics have demanded more attention to disaggregated studies at the group level (e.g. Cederman et al 2013). However, group level studies face a fundamental problem not encountered in the country/year set-up. While there is little disagreement on the number of countries in the world in any given year, there is no such agreement on the number of ethnic groups. Any sampling of the near infinite number of groups (if you permit dialects and sub-dialects to differentiate groups) will be arbitrary and subject to claims of bias. One goal of this paper is to present an approach that is sampling from one possible frame of ethnic groups, which will permit cumulative research on ethnic conflict based on disaggregated group-level data.

Empirical obstacles to examining the route to ethnic war

Established Selection Issues

In the study of ethnic conflict, selection issues are a recurring concern because the principal data used for empirical analysis, thus far, is based on the selection of groups that have already engaged with the state, as in the MAR data[6] and more recently groups that are "politically

[6] For a more extensive account of the selection bias problem as it pertains to the MAR data see Fearon and Laitin 2002, 2003; Fearon 2003; Birnir et al. 2015.

relevant" as in the Ethnic Power Relations (EPR) data (Wimmer, Cederman and Min 2009). Both data sets have been used to reveal patterns of conflict. But researchers need to be concerned about their sampling criterion and the conditions under which those patterns hold. Selection issues become especially problematic when we ask questions about what makes an ethnic group prone to violent conflict since both samples are selected on criteria that are likely to be correlated with a propensity for conflict. [7]

Selection bias is a fundamental problem for drawing either descriptive or causal inferences from data (Geddes; 2003; Shively, 2006; Hug 2010; Weidman, 2016). Selection biases are of many different types and cause distinct problems. One problem is that independent of concerns about estimating relationships between variables, much interest often centers on simple descriptive statistics about base rates in a population, which obviously cannot be estimated from a biased sample. It is likely, therefore, that we know less than we think about the prevalence of outcomes such as ethnic conflict.

A second selection concern focuses on detecting relationships between variables. When unrelated to the explanatory variable(s), selection on the dependent variable obscures a statistical effect where there really is one. For example, in the case of over selection of groups engaged in ethnic conflict, the reduction in variation on the dependent variable implies that we cannot, even if we had reasonable instruments, confidently determine the causes of rebellion using only the original MAR data. Without a representative sample of "ethnic groups" for each country, we cannot get confident estimates on the relationship between ethnic diversity and the frequency and type of ethnic conflict. Hug (2013) notes that it is likely that many "true" relationships go entirely undetected because of the aforementioned bias in the data. In particular he makes the

[7] See Vogt et al. 2015 for a discussion specifically pertaining to EPR.

case that contrary to the null finding of Gurr and Moore (1997), grievances likely do affect group propensity for rebellion.

A third selection concern is reporting bias.[8] Reporting bias happens in different ways. Weidman describes reporting bias in event data where the outcome is missed at random and some cases, therefore, get incorrectly coded. Random reporting errors likely impact estimates of relationships between variables as would noise. Alternatively, some outcomes are systematically more likely to be reported, sometimes as a function of explanatory variables (Weidman 2016). The principal concern with reporting bias centers on the latter types of cases where the reporting of the outcome is systematically related to purported explanatory variables. This type of bias may affect both the magnitude and direction of a correlation between an independent and a dependent variable (Hug 2010; Weidman, 2016).

Truncation of Group Data

A special class of selection concerns is the problem of truncated data.[9] In cases of truncated data, and unlike instances of selection on the dependent variable, values are included both where the outcome of interest occurs and does not occur. Furthermore, the dependent variable is not erroneously coded for some cases as in cases of reporting bias. In the case of truncation, what is especially worrisome with ethnic data lacking a coherent sample frame (Birnir et all 2015) is that data collection projects could be attuned to more obscure ethnic groups in one country but less so in another. Indeed, researchers have limited their selection of group

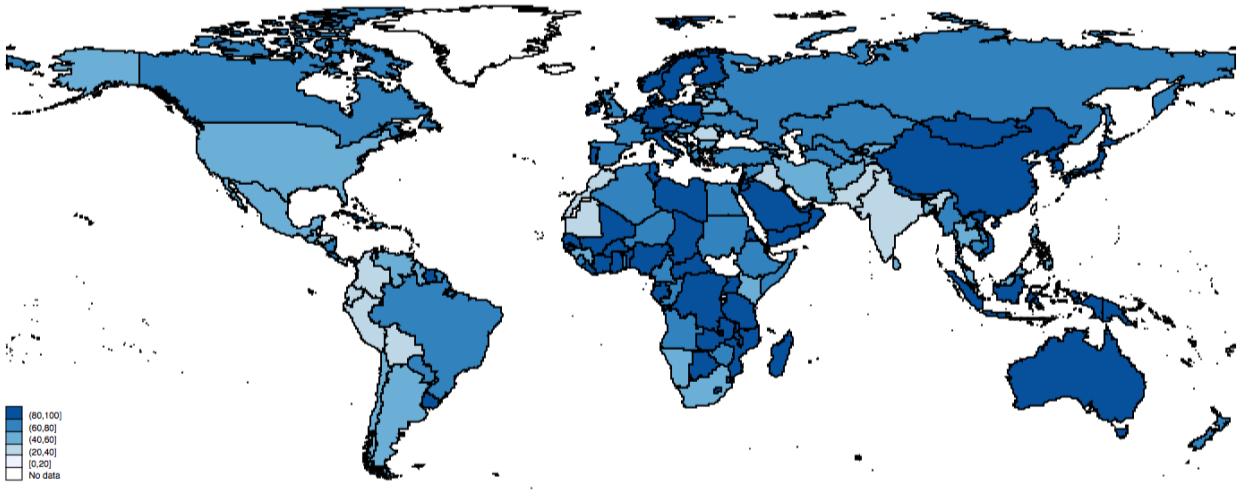
[8] Importantly, reporting bias presumes there exists a sample frame for coding of non-occurrences.

[9] The resultant problems in many cases likely resemble those occurring with reporting bias but the types of data suffering from each possibly differ.

level data by circumscribing the types of groups included – e.g., groups that are politically mobilized, discriminated against, or politically relevant – without estimating the implications of these limitations for their statistical estimations.

The effects of data truncation are of special interest in this paper because in data on ethnic groups this is likely a bigger problem than is reporting bias. Indeed, exploring group violence Fearon (2003) found little evidence of reporting bias in the MAR data – at least with respect to violent outcomes. Specifically, among the 539 groups not in MAR added by Fearon (2003), there were only 11 instances of rebellion between 1945 and 1998. [10] However, as shown in Figure 1, when comparing MAR with the AMAR sample frame where groups were selected irrespective of any political criteria (Birnir et al. 2015) a high percentage of socially relevant AMAR groups is missing from the original MAR data, especially in some of the most heterogeneous countries in the world.

Figure 1: By country: Proportion of Socially Relevant Groups in the AMAR Sample Frame but Missing From MAR.



A significant selection concern associated with truncated data, as with data suffering from

[10] This low reporting bias was independently confirmed by Brancati (personal conversation).

reporting bias, is that systematic truncation may render estimates of the relationship between the independent and dependent variable unreliable. To better understand how truncation affects these relationships, we constructed a generic simulation that systematically truncates data to drop more observations where the outcome did not occur – in ways that are also related to the explanatory variable. In short, we found that when compared to results from the "true" un-truncated data, coefficients estimating relationships between an independent and a dependent variable in the data that was systematically truncated varied substantially in size, sometimes even changing signs, when compared to the "true" correlation. Furthermore, we found standard errors to be invariably larger than in the "true" data, though this sometimes rendered correlations more significant and sometimes less, depending on the corresponding size of the biased coefficient. Judging by the simulation, truncation of data is, therefore, a significant threat to the accuracy of inference using uncorrected data on ethnic groups. (For details on the simulation see data Appendix).

Group Level Truncation in Country Level Analysis.

One of the potential problems resulting from truncation of group data (likely also a problem in data suffering from reporting bias) that has not been widely explored in the literature is error in inference when moving between levels of analysis. Despite receiving little attention, this type of error is possibly a serious problem in the literature because many studies use biased (by way of truncation) group level statistics to show correlations with a number of aggregate causal variables that do not vary within a country.

Specifically, the problem is that because of truncation, MAR and other datasets that select groups on some limited criteria provide an incorrect estimate of average group propensity to engage in outcomes such as violence, at levels more aggregated than the group, such as the country. If this limited ethnic group information is then regressed on measures that do not vary

within the country but only between countries, the resulting association is *not* necessarily an accurate indicator of ethnic group level propensity of engaging in the outcome of interest in a given country when compared to other countries. Instead, in many cases (at least where the total number of groups engaging in violence is high but the group proportion engaging in this activity is low) we will see positive correlations at the country level that henceforth have often been mistaken as indicators of group level propensity to engage in violence in any country.

This problem is best demonstrated with an example, as illustrated in Table 1. Suppose that in two countries X and Y there live 10 and 100 groups respectively. Hypothetical biased group-level data including information on all violent groups and some peaceful groups contains information on 8 groups from country X, 2 of which are violent and information about 20 groups from country Y, 10 of which are violent. The aggregate country level measure of group violence in countries X and Y would then show that 25% and 50% of groups engaging in violence respectively.

Suppose now that we were to collect information on the remaining two groups in country X and the remaining eighty groups in country Y, and find that the remaining groups in both countries are peaceful. Calculating the proportion of violent groups in each country we now find that 20% of all groups in country X engage in violence while only 10% of all groups in country Y ever engage in any violence. Consequently, *while it is still true that country Y experiences greater levels of violence than country X, it is also true that any one group in country Y is less likely to engage in violence than is any one group in country X.*

Table 1: Country vs. Group Propensity for Violence

Country	Aggregate measure of violence in a biased or incomplete sample of groups	Group measure of violence in a representative or complete sample of groups
X	25% (2/8)	20% (2/10)
Y	50% (10/20)	10% (10/100)

This problem is not commonly discussed in the literature on ethnic conflict that often uses biased group data to make inferences about group propensities.[11] Consider examples of country level measures that in the literature have been associated with group propensity to engage in violence. These include ethnic fractionalization measures (Reagan and Norton 2005; Olzak 2006), measures of political institutions (Saideman et al 2002; Alonso and Ruiz-Rufino 2007), and measures of country level development (Cetinyan 2002; Walter 2006). While inferences have been made from these measures for group level measures of violence, in all likelihood these studies are really measuring country level probabilities.

All MAR (AMAR) Sample Frame.

Weidman (2016) laments that current probes of selection bias issues alternately assume away the problem; only focus on sensitivity analyses assuming the direction of bias without any evidence; or rely on estimators that require strong statistical assumptions. Instead, he suggests that whenever possible, using real data to establish and solve the problem is a preferable solution.

The sensitivity analyses in the simulation described above confirmed our suspicion that truncation of data likely presents a problem for analytical inference. Heeding Weidman's call, we

[11] The problem of truncation is also separate from the problem of reporting bias because even in the instances of substantial reporting bias the average values between more aggregate units may still maintain their relative order.

outline below our solution involving the collection of real sample data from the AMAR sample frame of socially relevant groups (Birnir et al 2015). Describing the construction of the AMAR selection frame of "socially relevant ethnic groups," Birnir et al. (2015) follow Fearon (2006) in defining socially relevant as "when people notice and condition their actions on ethnic distinctions in everyday life." Social (and political) identities, in turn, are subsets of all existing ethnic structures. [12] Importantly, social relevance of an identity does not refer to political mobilization and does not have inherent political connotations but only refers to the salience of the identity in guiding an individual's actions in her life.

In our proposed solution, we acknowledge that no single list of ethnic groups is the correct one for every context. Furthermore, as noted by Fearon (2003) in coding ethnic groups, "It rapidly becomes clear that one must make all manner of borderline- arbitrary decisions, and that in many cases there simply is no single right answer to the question 'What are the ethnic groups in this country?' Constructivist or instrumentalist arguments about the contingent, fuzzy, and situational character of ethnicity seem amply supported" (2003:197).

Therefore, to probe the usefulness of the AMAR sample frame for our purposes we first compared it to multiple other data collection efforts. Specifically, we matched the AMAR sample frame by country and group name[13] to groups in: Fearon 2003, Alesina et al.2003 and EPR 2010 (version 1.1) and 2012 (version 3). The results of our comparison demonstrated in Table 2

[12] Chandra and Wilkinson define ethnic structure as "distribution of descent-based attributes— and, therefore, the sets of nominal identities— that all individuals in a population possess, whether they identify with them or not." (2008:523)

[13] We did not verify geographic overlap, as the data do not typically contain geographic indicators for group location within a country.

suggest that there is substantial agreement about the configuration of "socially relevant" ethnic groups at the national level across datasets collected independently and using very different coding rules. The principal differences we found between these lists were due to differences in project definitions and/or objective, in aggregation, and in inclusion parameters. Despite these differences, we found that there was great overlap among the lists. Nearly all of the groups enumerated in these other lists were either in the main AMAR list or listed as sub-groups of AMAR umbrella groups (for a suggestive list of subgroups see Birnir et al. 2015).

Even though there is significant overlap between AMAR and these existing ethnic group data, due to differences in selection criteria the AMAR data contain substantially more ethnic groups. As seen in Table 2, in the case of Alesina et al., AMAR contains 477 more groups, in the case of Fearon, 410 more groups, and in the case of EPR v1.1 and v3.0, 502 and 500 more groups. Table 2 thus reveals that a loose set of commonly identified "socially relevant" groups is shared across datasets, even if the notion of an ethnic group is fuzzy. (For further discussion relating to group overlap see the data Appendix).

Table 2: comparing ethnic groups across Alesina, EPR, Fearon[14] and AMAR.[15]

		Alesina	EPR v1.1 2010[16]	EPR v3.0 2012	Fearon
TOTAL IN LISTS	**Total number of groups (AMAR 1202)**	1054	731	758	858
TOTAL MATCHED	**Total number matched with AMAR**	693 66%	656 90%	677 89%	778 91%
Thereof	**Full match (full congruence)**	505 73%	513 78%	524 77%	637 82%
	Match, but at different level of aggregation (group is match to AMAR sub-group of aggregate group)	98 14%	98 15%	104 15%	65 8%

[14] The total number of groups in the Fearon 2003 paper is 822. We received an updated version of the 2003 data from James Fearon, which consists of 858 groups; this version of the data was used for the AMAR match.

[15] The total number of groups in AMAR is 1202. Of those 288 are current MAR groups that account for 291 AMAR groups. See fn. 16.

[16] Per EPR, the 1.1 version of the data "includes annual data on over 733 groups" (http://www.icr.ethz.ch/data/other/epr_old). However, after downloading the MASTER_EPR_v1.1 version of the data from the Harvard Dataverse (http://thedata.harvard.edu/dvn/dv/epr/faces/study/StudyPage.xhtml?globalId=hdl:1902.1/11796&tab=files&studyListingIndex=0_a777931694382ee99a6e0f5576cb) and removing duplicate groups based on the cowgroupid variable, we found 731 unique groups. We also downloaded and removed duplicate entries for the EPR_groupyear_v1.1 version of the data, and found 728 unique groups (this version of the data is missing the Kpelle and Kru in Liberia and the Northern Hill Tribes in Thailand, which appear in the MASTER version of the data). Therefore, the total number of groups evaluated for this match is 731, and not 733.

	Match, but group combined with another group into one aggregate in AMAR	45 7%	6 1%	6 1%	40 5%
	Match, but group is listed as two or more groups in AMAR	45 6%	39 6%	43 6%	35 4%
TOTAL NOT MATCHED	**Total number not matched**	361 34%	75 10%	81 11%	80 9%
Thereof	**Not matched because don't meet AMAR population threshold criteria**	80 24%	66 88%	67 83%	22 28%
	Not matched because don't meet AMAR ethnic criteria	6 2%	2 3%	3 4%	3 4%
	Not matched because lack of available data	4 1%	0 0%	0 0%	0 0%
	Not matched because countries don't meet AMAR population threshold, or former communist states not coded in AMAR[17]	133 37%	6 8%	11 14%	55 69%
	Not matched because group names were not provided or group was in "other category"[18]	129 36%	1 1%	0 0%	0 0%
	Not matched because coded in another country or for other similar reasons	9 2%	0 0%	0 0%	0 0%
TOTAL MATCHED + NOT MATCHED	**Total AMAR matched + AMAR not matched**	1054 100%	731 100%	758 100%	858 100%
TOTAL IN AMAR ONLY	**Total groups in AMAR only**	477	502	500	410

[17] The AMAR lists groups in current nation states only.

[18] If Alesina et al. (2003) or EPR subsumed groups in a category called “other” or omitted the group name we could not match those groups with named AMAR groups.

Before turning to the sample construction, a few more words on validity are in order. The AMAR list includes groups over a certain population threshold and at a certain level of aggregation. Therefore, the groups counted are not a comprehensive list of all possible socially relevant ethnic groups but a reasonable minimum list of sizeable and nationally recognizable groups. In our review of the list we became aware of several other groups that were right at the population boundary for inclusion or whose population numbers we could not verify for inclusion. Therefore, while prior work suggests it is unlikely that AMAR has over-counted peaceful groups, it is possible that at the margins AMAR undercounts peaceful groups because of errors in aggregating groups that should be left disaggregated or exclusion of groups that really should be included separately in the sample frame. The implications of such errors for our estimates of the frequency of conflict would be to increase the weight of peaceful groups. Proportionally this undercounting would, therefore, make the instances of ethnic rebellion against the state even less common, re-introducing the bias that previous work has suffered from, though far less acutely. This suggests there is good reason to continue refining the sample frame but raises no flags for the suggestive analysis that follows.

The AMAR Sample of Socially Relevant Groups

Confident that there is broad scholarly consensus about the core of the frame of socially relevant ethnic groups – of which AMAR is the most extensive enumeration - we can now describe how we constructed our AMAR sample and how the fully coded sample data can be used in analysis.

The foundation of our sample is the original MAR dataset. As noted earlier, MAR-listed groups, though only a fraction of all ethnic groups, have been implicated in nearly all instances of ethnic rebellion against the state. But to construct a representative sample of ethnic

groups that includes both violent groups that are likely adequately represented in existing ethnic group data and peaceful groups that are likely underrepresented in current ethnic group data, for the AMAR sample we needed to go beyond the 288 MAR groups that are continuously coded from the time they enter the dataset. Using the list of socially relevant ethnic groups introduced in Birnir et al. 2015, we consequently add a random stratified sample from the roughly 900 NEW (not in the original MAR) AMAR groups that meet the MAR criteria, listed below, for inclusion:

1. Membership in the group is determined primarily by descent by both members and non-members. (The group may be a caste determined by descent.)
2. Membership in the group is recognized and viewed as important by members and/or non-members. The importance may be psychological, normative, and/or strategic.
3. Members share some distinguishing cultural features, such as common language, religion, occupational niche, and customs.
4. One or more of these cultural features are either practiced by a majority of the group or preserved and studied by a set of members who are broadly respected by the wider membership for so doing.
5. The group has at least 100,000 members or constitutes one percent of a country's population.

Together these observations (MAR and a stratified random sample of NEW) constitute the AMAR sample introduced in this paper.

The MAR portion of the sample:

The original MAR data was coded in four distinct phases.[19] Researchers working with the data at any given time likely use the current cases only. Therefore, we also focus here on the core of 288 cases that are current (in the AMAR sample frame these appear as 291 cases as some

[19] For a detailed history of coding phases as relevant to this project please see the AMAR codebook. Earlier versions of this paper compared MAR cases coded at any stage to cases listed as NEW. This could, however, lead to double counting of cases that exited MAR and entered again as NEW.

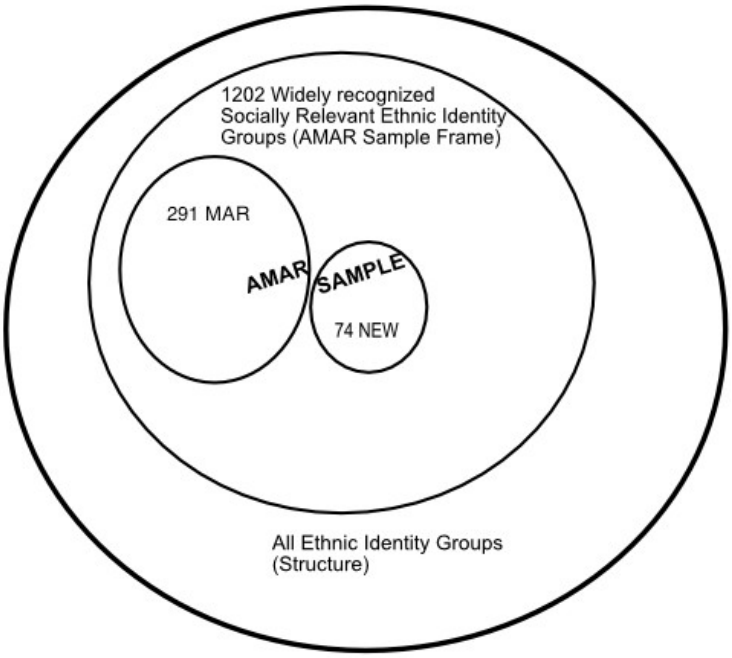
original MAR groups were split, merged or dropped in the AMAR enumeration – for a list of these groups see data Appendix). As shown in Figure 1 the original MAR data disproportionately lack groups from certain regions as judged by comparison with the AMAR list. Additional comparisons not included here show that the MAR data also systematically miss groups of a particular size. Consequently, we also correct for this regional and population under-sampling in our analysis.

The NEW portion of the sample:

So as not to replicate in our sampling of NEW the known regional and population biases in MAR, we used a three-tiered population strata of small, medium and large groups in each region: small groups are groups whose population constitutes less than or equal to 2 percent of a county's population; medium groups account for over 2 percent but less than or equal to 20 percent of a country's population; and large groups number more than 20 percent of a country's population. Using this stratification we generated a list of 100 groups from NEW, of which a random 74 were coded for all the principal variables of MAR. [20] See Figure 2 that illustrates inter alia where our 74 groups fit into the overall ethnic structure, and into the AMAR sample frame.

[20] Resources available have permitted only these 74 of the 100 cases to be coded to date; the 74 are a random sample of the 100.

Figure 2: Universe of ethnic structure, AMAR list of socially relevant and widely recognized ethnic groups, and the AMAR sample.



Sample Correction

When two sample-segments are analyzed together to produce descriptive or inferential statistics for the population parameter of interest, each segment (here MAR and NEW, alternatively males and females or any other category) can be assigned weights according to their relative importance in the population, when a pertinent sample frame is available. Weighting is common in survey analysis, where sample segments often over- or under-represent particular population segments (Kalton, G. 1983; Kalton and Flores-Cervantes, I. 2003; National Election Study; Chromy and Abeyasekera 2005; Stoop, Billiet, Koch, and Fitzgerald 2010). We follow the common strategy of defining weights as the inversed sampling probability of an individual observation. Addressing concerns about large variability in weighting adjustments inflating the variance in survey responses (here sample segment values) (Kalton and Flores 2003) we

conservatively scale our weights for the size of the sample and sample frame (see data Appendix for detailed description of weights).

The obvious concern is how well the weighted sample captures reality, assuming the sample frame is reasonably accurate. To explore this issue we collected the group's numerical proportion of the population for the entire AMAR sample frame (i.e. the entire set of 1202 groups). Thus we can calculate the true average group proportion of the population in the entire sample frame. Prior to calculating this average, we exclude politically dominant groups where there is a single dominant group in a country because we do not want to analyze the correlates of violence for ethnic groups that control the government and are considered to be the state. [21] Excluding politically dominant groups where there is a single dominant group in a country (which leaves 1085 groups[22]), the average population proportion of socially relevant groups is 6.70 percent. In the MAR data the average group proportion of the population excluding dominant groups in countries where there is a single dominant group is 11.17 percent. In contrast, the weighted estimate of minority group proportion in the AMAR sample is 7.15 percent, which approximates the full set of cases.

[21] Politically dominant refers to a group that consistently controls or is a senior partner in the executive in democratic countries or the equivalent in authoritarian countries. We used EPR coding for this information supplementing with country specific accounts where EPR did not code these data.

[22] So as not to throw away data the actual number we use in the analysis is 1089 because there are 4 "extra" groups in the MAR data that don't count in the sample frame because they were split into more than one group in the AMAR sample frame. We account for this difference in the assignment of the weights.

The Data.

In sum, the data presented in this paper consists of three parts. The first is the integrated and cleaned classic MAR data. The second is the 74 NEW groups randomly selected from the remaining AMAR sample frame (NEW), fully coded for all current MAR variables and integrated with the original MAR data along with inverse probability weights for each sample segment. Third is the coding of classificatory variables and analytical variables for the entire AMAR sample. We discuss each in turn below.

MAR integration

In 2006, after a review of the approximately 400 variables that had been part of the various phases of the MAR project since its inception, a total of 71 variables were selected as being "core" variables for Phase V of the data collection. Of the "core" variables, some were then reformulated to facilitate either collection or statistical analysis of the data. Specifically, there were two types of reformulations: variable levels were re-specified from Phase IV to Phase V (i.e., levels 1, 2, 3 in Phase IV were reformulated to 0, 1, 2 in Phase V) and variables were changed from Phase IV to Phase V (i.e., political grievances were coded with 12 variables in Phase IV but were reformulated into only 1 variable for Phase V). While this review and reformulation brought the data in line with current research interests in the field, the variables that were reformulated for Phase V data were not reconciled with earlier phases. Consequently, for the reformulated variables the two parts of the data (before and after 2004) could not be used together. As a part of the AMAR project, we reconciled most of the reformulated variables from the various phases of the MAR data collection into one dataset creating variables that are

continuously coded across the distinct phases of the project. [23]

In addition, as a part of AMAR Phase I we systematized and integrated extant community input into the data. The community input we incorporated is of two kinds. The first consists of recode requests documented over many years and the second are discrete variables that were coded and/or updated by scholars who were intimately familiar with the project and undertook independent data collections in line with MAR protocols. All of this work is detailed in the new codebook accompanying the data, including also the code for reconciliation of the various MAR Phases.

The 74 NEW AMAR Groups Coded for All Core MAR Variables.

As noted, 74 groups out of the over 900 NEW AMAR groups – not previously in MAR – were randomly selected and coded annually from 1980 to 2006 for all core variables in the MAR data. These groups are listed in Table 3.

[23] In total, 39 variables were reformulated in Phase V of the MAR data. Of these, 21 variables were reformulated for the AMAR Phase I data. The remaining variables required substantial research in order to be reconciled and could not be completed for the AMAR Phase I data due to funding constraints. For a detailed description of these variables and the reformulation process, see the codebook.

Table 3: 74 New Groups Coded for all MAR variables.

Country	Group
Afghanistan	Aimaq
Afghanistan	Brahui
Angola	Nyaneka-Humbe
Bangladesh	Garo
Bangladesh	Santals
Botswana	Kgalagadi
Burkina Faso	Busansi
Burkina Faso	Songhay
Burma	Lahu
Cambodia	Chinese
Chile	White/Mestizo
China	Mongolian
Colombia	Mestizo/White
Democratic Republic of Congo	Chokwe/Koko/Tshokwe
Democratic Republic of Congo	Logo/Logokuli
Democratic Republic of Congo	Lugbara
Democratic Republic of Congo	Bemba/Shila
Gabon	Kota
Gambia	Fulani
India	Scheduled Tribes Of East India
India	Syrian/Malabar Christians
Indonesia	Pasemah
Jamaica	Mixed
Japan	Filipinos
Kazakhstan	Tatar/Tartar
Kazakhstan	Uighur
Kenya	Turkana
Latvia	Lithuanian
Latvia	Polish
Lebanon	Armenian
Macedonia	Turks
Malaysia	Orang Asli
Mali	Fulani/Fulbe/Peuls
Mali	Mande
Mali	Maures/Moors
Mozambique	Makonde
Namibia	Ovambo
Nepal	Kirata/Kiranti/Kirati

Nepal	Sherpa
Netherlands	Frisians
New Zealand	Asians
Nigeria	Kamberi
Nigeria	Plateau Chadic
Pakistan	Seraiki/Saraiki
Papua New Guinea	Kamano
Peru	Asians
Russia	Dargins
Russia	Kalmyks
Russia	Komi
Saudi Arabia	Egyptians
Serbia	Serbs
Somalia	Bantu (Non-Somali)
Spain	Valencian
Sri Lanka	Sinhalese
Sudan	Arab/Ja'Aliyin
Swaziland	Zulu
Syria	Druze
Tanzania	Gogo
Tanzania	Iraqw, Mbulu
Tanzania	Luo
Thailand	Thai
Timor-Leste	Papuan
Tunisia	Berber
Uganda	Banyoro
Uganda	Nkole/Nkore
Uganda	South Asians
United Kingdom	Welsh
Uruguay	White/Mestizo
Vietnam	Hmong
Yemen	Sunni Arabs
Zambia	Tonga-Ila-Lenje
Zimbabwe	Kunda/Seba
Zimbabwe	Lozi
Zimbabwe	Nyanja

The coded core variables, listed in Table 4, are of four types and fully described in the AMAR Phase I codebook. The first category is a suite of group characteristics including group identity and group concentration. The second category is a group status suite including variables accounting for autonomy and group grievances. The third suite of variables accounts for external support by state and non-state actors. The fourth suite accounts for group conflict behavior and state repression.

Table 4: Core MAR Variables Coded for the 74 New Groups.

Variable Name	Variable Description
NUMCODE	Ethnic group case identifier (country code + group id)
CCCODE	Country ID number (The Correlates of War (Singer and Small) country identification number
COUNTRY	Country in which the group resides
REGION	AMAR regions
YEAR	Year of Observation
GPOP	Group's population
CPOP	Country's population
GPRO	Group proportion of country population
LANG	Different language group
CUSTOM	Different group customs (marriage, family, dress, etc.)
BELIEF	Different group religion
RELIGS1	Specific religion: Plurality religion of group
RACE	Different physical appearance
GROUPCON	Group spatial distribution
GC119	Urban/rural distribution
GC2	Regional base
GC6B	Regional base--proportion of group members in regional base
GC7	Proportion of group living outside regional base
GC10	Transnational dispersion -- kindred groups
GC11	Transnational dispersion -- kindred groups in power
AUTLOST	Index of lost political autonomy, based on YEARWT, MAGN, PRSTAT
YEARWT	Year of the most recent less of autonomy
MAGN	Magnitude of change
PRSTAT	Group status prior to change
AUTONEND	Year/decade/century autonomy was lost
TRANSYR	Year/decade/century transferred
SEPX	Separatism index
SEPKIN	Active separatism among kin groups
EMIG	Emigration for political or economic reasons
DISPLACE	Internal displacement for political or economic reasons
POLDIS	Political discrimination index
ECDIS	Economic discrimination index
CULPO1	Restrictions on religion
CULPO2	Restrictions on use of language or language instruction
GOJPA	Group organization for joint political action
AUTON2	Group autonomy status
AUTGAIN	Year group gained autonomy
AUTPRO	Percentage of group in autonomous region
LEGISREP	Group representation in legislative branch of central government

EXECREP	Group representation in executive branch of central government
GUARREP	Group is guaranteed representation in central government
POLGR	Highest level of political grievance
ECGR	Highest level of economic grievance
CULGR	Highest level of cultural grievance
KINSUP	Any kindred group support
KINMATSUP	Kindred group material, non-military, support
KINPOLSUP	Kindred group political support
KINMILSUP	Kindred group military support
STASUP	Any foreign state or IGO support
STAMATSUP	Foreign state/IGO material, non-military, support
STAPOLSUP	Foreign state/IGO political support
STAMILSUP	Foreign state/IGO military support
NSASUP	Any non-state actor support
NSAMATSUP	Non-state actor material, non-military, support
NSAPOLSUP	Non-state actor political support
NSAMILSUP	Non-state actor military support
INTRACON	Presence of intracommunal conflict
FACTCC1	Names of intracommunal antagonists with highest level of conflict
FACTSEV1	Severity of conflict for first pair of antagonists
FACTCC2	Names of intracommunal antagonists with second-highest level of conflict
FACTSEV2	Severity of conflict for second pair of antagonists
FACTCC3	Names of intracommunal antagonists with third-highest level of conflict
FACTSEV3	Severity of conflict for third pair of antagonists
INTERCON	Presence of intercommunal conflict
CCGROUP1	Name of group with highest level of conflict
CCGROUPSEV1	Level of conflict with CCGROUP1
CCGROUP2	Name of group with second-highest level of conflict
CCGROUPSEV2	Level of conflict with CCGROUP2
CCGROUP3	Name of group with second-highest level of conflict
CCGROUPSEV3	Level of conflict with CCGROUP3
PROT	Protest
REB	Rebellion
REPGENCIV	Repression of group civilian populations (those not engaging in violent or nonviolent political activities)
REPNVIOL	Repression of group members engaged in nonviolent collective action
REPVIOL	Repression of group members engaged in violent collective action

AMAR Sample Frame Variables

The final data contribution of this project consists of new AMAR variables that we coded for the entire AMAR sample frame of 1202 groups. These are of three kinds as noted in Table 5. The first set identifies the group and consists of variables already present in the MAR data that were expanded to account for all AMAR groups. The second set of variables is classificatory. These variables detail whether the case is an ethnic group that was originally in the MAR data or belongs to the set of NEW groups in the AMAR sample frame. Furthermore, a series of variables account for whether the original MAR cases changed in any way between the MAR data and the AMAR sample frame, either by splitting the group or merging with another group or changing the name of the group. Finally, a separate variable accounts for the 74 randomly selected groups that were fully coded for MAR variables.

The third set of AMAR variables is functional. The first of these is group proportion of national population for every group in the AMAR sample frame. A second set of variables accounts for whether the group has at any point been politically dominant. A third set of variables accounts for whether the AMAR sample frame group has a match in Fearon (2003), Alesina et al. (2003), or in either of the 2 matched versions of EPR. Detailed information on all of these variables is included in the new AMAR Phase I codebook.

Table 5. List of Variables Coded for all 1202 AMAR Groups.

Variable Name	Variable Description
NUMCODE	Ethnic group case identifier (country code + group id)
AMAR GROUP	Full name of AMAR ethnic group
CCCODE	Country ID number (The Correlates of War (Singer and Small) country identification number
COUNTRY	Country in which the group resides
MAR PROPER	NEW TO AMAR - Group coded in MAR proper data (MAR Phase I-V)
SELECTION BIAS	NEW TO AMAR - Group coded in AMAR selection bias data
NAME CHANGED	NEW TO AMAR - Group's name changed from MAR to AMAR
PREVIOUS NAME	NEW TO AMAR - Name of group as appeared in MAR proper
SPLIT GROUP	NEW TO AMAR - Group split from MAR to AMAR
MERGED GROUP	NEW TO AMAR - Group merged from MAR to AMAR
ONE DOM GROUP	NEW TO AMAR - One politically dominant group in the country
ALL DOM GROUPS	NEW TO AMAR - Politically dominant groups
ALESINA MATCH	NEW TO AMAR - AMAR group matched to ethnic group in Alesina et al. 2003 data
EPR V1 MATCH	NEW TO AMAR - AMAR group matched to ethnic group in Ethnic Power Relations (EPR) v1 2010 data
EPR V3 MATCH	NEW TO AMAR - AMAR group matched to ethnic group in Ethnic Power Relations (EPR) v1 2010 data
FEARON MATCH	NEW TO AMAR - AMAR group matched to ethnic group in Fearon 2003 data
GPRO AMAR	NEW TO AMAR - Group proportion of country population for all AMAR groups

Future Directions for Research: Frequency and Causes of Ethnic Rebellion

Until now we have not had a good idea of either the frequency of ethnic violence against the state or of its causes because we lacked a representative sample of ethnic groups to study. Having such a coded sample we are now able to demonstrate descriptively that previous answers have been systematically biased. Indeed, the problem of truncation in ethnic data outlined earlier and the simulation showing the sensitivity of correlation coefficients and standard errors in truncated data indicate there is good reason to revisit many of the purported correlates of ethnic

politics – especially in national level correlations. Here we suggest some directions forward, hopefully to motivate future research relying on our proposed frame.

Frequency of ethnic rebellion.

Table 6 compares the frequency of ethnic minority violence as recorded in the uncorrected MAR sample (first column) to the corrected weighted AMAR sample (second column). The variables listed are any rebellion against the state (coded as 1 if the group has engaged in any rebellion against the state since 1945 or 1980, and 0 otherwise); high levels of rebellion against the state (coded as 1 if the group has engaged in small scale guerilla activity with a coding of 4 or greater for level of rebellion since 1945 or 1980, and 0 otherwise); the average level of group violence since 1945 or since 1980; and the highest level of ethnic violence by the group since 1945 or since 1980. [24]

The more representative sample, excluding politically dominant groups where there is a single dominant group in a country (again, we exclude these groups because we do not want to calculate correlates for ethnic groups that control the government and are considered to be the state; see fn 20), shows that at the ethnic group level, ethnic rebellion against the state is far rarer than what one would infer from the MAR data. Specifically, when coded as a binary variable to account for any instances of rebellion (including all types of rebellion from the lowest level to ethnic war), the MAR data suggest that two thirds of all minority groups have at some point since

[24] In MAR and AMAR rebellion is coded on an ordinal 7 point scale in addition to a coding of 0 when no rebellion is reported. A rebellion code of 1 indicates political banditry, 2 campaigns of terrorism, 3 local rebellions, 4 small-scale guerrilla activity, 5 intermediate guerrilla activity, 6 large-scale guerrilla activity, and 7 civil war. The code used to indicate there is no basis for judgment is -99.

1945 engaged in violence against the state. In contrast, the weighted average rebellion in the AMAR sample suggests that this number is far lower, at 29% of all widely recognized groups (majorities and minorities) having ever engaged in rebellion against the state. The MAR data suggests that well over a third of all ethnic groups have engaged in high levels of ethnic rebellion whereas the corrected AMAR data suggests that number is below 17%. The MAR data suggests that the average magnitude of rebellion is low with the vast majority of groups that do engage in rebellion only perpetrating very low levels of violence. The AMAR sample suggests that the average magnitude of rebellion is lower still. The same is true for average group maximum levels of rebellion: the AMAR sample averages are less than half of MAR averages.

Table 6: Comparing the share of groups that have engaged in violence against the state MAR and the weighted AMAR sample.

Variable	Group Violence against the state MAR only	Group Violence against the state AMAR weighted sample
Proportion of groups engaging in any rebellion since 1945	0.617	0.285
Proportion of groups engaging in any rebellion since 1980	0.579	0.274
Proportion of groups engaging in high levels of rebellion (4 or higher) since 1945	0.354	0.167
Proportion of groups engaging in high levels of rebellion (4 or higher) since 1980	0.313	0.158
Average magnitude of rebellion since 1945	0.818	0.259
Average magnitude of rebellion since 1980	0.837	0.268
Average maximum level of rebellion since 1945	2.645	1.147
Average maximum level of rebellion since 1980	2.328	1.070

In sum, the frequency of the combativeness of ethnic groups is greatly exaggerated as a result of selection issues – notably, the previously underexplored issue of truncation - in the uncorrected dataset. Differences in the average magnitude of rebellion are even farther off the target when comparing MAR to the AMAR weighted sample. While many scholars assumed this to be the case, our AMAR dataset allows us precisely to estimate the degree of bias in past reckonings.

Correlates of Ethnic Rebellion: Group Level Data

Turning now to the correlates of group rebellion, with the truncated ethnic group data we don't know the implications of limited selection. It could simply decrease variance in the

dependent variable in ways that make it more difficult to detect relationships with explanatory variables unrelated to the selection criteria. Or if the selection on the dependent variable is systematically related to explanatory variables, this would render suspect the purported relationships. To examine both possibilities, Table 7 compares correlations between the biased MAR group level data and the corrected weighted AMAR sample for some commonly examined group level correlates of rebellion. Specifically, we ran bivariate regressions on cross- sectional group level data, with standard errors clustered at the country level. The regressions correlate political, economic, and cultural grievances and a measure of group concentration to various measures of violence. Table 7 substantiates both the concern that researchers have incorrectly estimated the magnitude of relationships and missed important correlations. Nearly all of the associations show a substantial difference in the magnitudes of the effects estimated with the biased MAR sample as compared with the weighted AMAR sample when group concentration, political, economic and cultural grievances are correlated with a set of rebellion measures. Furthermore, in most of the MAR sample neither economic grievances nor cultural grievances are significantly correlated with outcomes whereas the corrected AMAR sample demonstrates a significant correlation with both economic and cultural grievances for all the rebellion measures. Group concentration, in turn, is not as clearly a substantial driver of conflict in the corrected sample as the original MAR data would suggest. While the robustness of these suggested new relationships need to be tested more rigorously, the concern about selection bias suppressing relationships between variables as described by Geddes and Shively and possibly distorting other relationships as emphasized by Hug and Weidman and others is well founded in the truncated MAR data.

Table 7: Re-estimating group level correlates of civil war onsets. Bivariate regressions, Standard Errors clustered on country.

	MAR only				**AMAR weighted sample***			
Variable	Any Rebellion since 1945	Average magnitude of Rebellion since 1945	Max. Rebellion since 1945	Rebellion 4 or over since 1945	Any Rebellion since 1945	Average magnitude of Rebellion since 1945	Max. Rebellion since 1945	Rebellion 4 or over since 1945
Group con-centration	0.162** (0.027)	0.343** (0.073)	0.828** (0.174)	0.134** (0.028)	0.060 (0.039)	0.078* (0.038)	0.224 (0.222)	0.042 (0.038)
Political grievances	0.216** (0.027)	0.613** (0.098)	1.131** (0.183)	0.181** (0.034)	0.255** (0.030)	0.360** (0.063)	0.996** (0.171)	0.145** (0.032)
Economic grievances	0.044 (0.041)	0.232* (0.102)	0.166 (0.242)	0.060 (0.044)	0.186** (0.003)	0.292** (0.066)	0.710** (0.217)	0.090* (0.035)
Cultural grievances	-0.002 (0.053)	0.007 (0.126)	-0.258 (0.320)	-0.017 (0.055)	0.172** (0.053)	0.254** (0.066)	0.592** (0.222)	0.059# (0.033)

Standard Errors in Brackets (#p<.10, *p<.05, **p<.01)

Ethnic Rebellion: Country and Group Level Data

A related but less explored concern raised in this paper is that of erroneous inferences from country level data correlated with truncated group level data as if the truncated data correctly represented country group averages. One such debated relationship is the correlation between ethnic heterogeneity and minority rebellion against the state.[25] Ethnic diversity is often considered a country-level attribute associated with violence and state deterioration (Chandra 2012).[26] Looking at the correlations in Table 5 from the simple bivariate regressions, with

[25] Not only is there a scholarly consensus on what constitutes a "socially relevant" ethnic group, but the Ethnolinguistic Fractionalization (ELF) scores by country across the datasets are also nearly alike. For example, the AMAR measure correlates at .78 with Alesina's measure and .89 with Fearon's measure.

[26] For exceptions in the literature on civil war see Fearon and Laitin (2003) and the subsequent

standard errors clustered on country, between the measure of Ethnolinguistic Fractionalization (ELF – static at the country level) and rebellion for MAR groups only, it is easy to see why the literature thus far associates ethnic heterogeneity with violence. The first column of Table 8 accounts for the substantial and significant association in the MAR data between ELF and every indicator of rebellion. In contrast, the second column correlates minority violence against the state with ELF in the weighted AMAR sample. The magnitudes of every coefficient are drastically reduced and none are significant.

Table 8: Comparing the bivariate association between Ethnic Fractionalization (ELF) and violence in MAR and the weighted AMAR sample.

Variable	The correlation between ethnic heterogeneity (ELF) and violence MAR only	The correlation between ethnic heterogeneity (ELF) and violence AMAR sample weighted
Any rebellion since 1945	0.407**	0.027
	(0.141)	(0.133)
Average magnitude of rebellion since 1945	0.724#	0.081
	(0.396)	(0.140)
Maximum level of rebellion since 1945	2.687**	0.794
	(0.974)	(0.565)
Rebellion level 4 or greater since 1945	0.403*	0.10
	(0.164)	(0.086)

Standard Errors in Brackets (#p<.10, *p<.05, **p<.01)

Assertions about the detrimental effect of ethnic heterogeneity on group propensity to commit violence are, if these relationship holds up to econometric scrutiny, substantially exaggerated. There is no reason to believe that groups in heterogeneous societies are more likely than groups in homogeneous societies to engage in violence or that their average violence is

---

literature on civil war onset. In the literature on minority rebellion against the state see Birnir 2007 and Cederman et al 2010 for exceptions.

higher. The reason for the difference between the results obtained with the un-weighted MAR data and the weighted AMAR sample is likely that the truncation of the MAR data (or any other truncated group level data) is systematically related to heterogeneity. Where there are more groups, more are on average missed, especially if they are peaceful. Hence, the positive correlation between heterogeneity and conflict is likely correct at the country level. At the group level, however, violence is not more likely because the high likelihood of violence in the country is not a good predictor of any particular group involvement in violence, especially when there are many groups.

Another way to think about this is that rebellious groups are more likely to come from ethnically heterogeneous countries if only because there are more groups to count and more to miss. Importantly, because rebellious groups (that tend to come from heterogeneous countries) are overrepresented in MAR, the first set of correlations estimates the likelihood of rebellion occurring in a given country rather than the effect of heterogeneity on the likelihood of any group engaging in rebellion. When we add in the NEW AMAR data (in which there are more peaceful groups – often in high conflict heterogeneous countries) and weight the data to count these NEW AMAR groups in proportion to their weight in the sample frame, the correlation between fragmentation, rebellion onset and average rebellion that appears present in MAR vanishes for all of our indicators.

Another well-known conundrum in the literature is the apparent relationship between development and violence. Because rebellion is relatively more common in the developing world, some hold that poverty causes rebellion. A common retort is that groups need resources to rebel, thus poor groups should be less likely to engage in violence. We suspect that both perspectives are right. Poverty is likely a grievance but one that only a few groups can act upon.

Thus, developing countries should experience greater levels of violence overall (if only due to state weakness in deterring rebellion) but fewer groups in any given developing country should have the resources to rebel.

Consequently, we would expect a country level association between poverty and aggregate measures of rebellion. However, when accounting for the probability that any particular group will rebel we would expect this association to be reduced. To examine this expectation, Table 9 shows Penn World measures of logged GDP per capita correlated (through bivariate regressions clustered on country) with a number of measures of rebellion: viz., whether a group has ever engaged in rebellion, average levels of rebellion, maximum levels of group rebellion and rebellion by groups engaged in high level conflict only.

Table 9: Comparing the bivariate associations between the log of GDP per capita and violence in MAR and the weighted AMAR sample.

Variable	The correlation between log of GDP per capita and violence MAR only*	The correlation between log of GDP per capita and violence AMAR sample weighted*
Any rebellion since 1945	-0.077** (0.026)	-0.010 (0.027)
Average magnitude of rebellion since 1945	-0.277** (0.072)	-0.044 (0.030)
Maximum level of rebellion since 1945	-0.839** (0.161)	-0.288** (0.105)
Rebellion level 4 or Greater	-0.148** (0.026)	-0.051** (0.016)

Standard Errors in Brackets (#p<.10, *p<.05, **p<.01)

As expected, the more limited MAR data – likely accounting for country level characteristics – suggests that occurrence of ethnic rebellion, average ethnic rebellion and maximum levels of violence are all negatively related to a country's wealth. Overall poorer countries appear more likely to experience rebellion. In contrast, and also as expected, the group

level AMAR data suggests that very few poor groups have the opportunity to act upon their grievances. Thus poorer groups seem no more likely to be embroiled in a rebellion than are their counterparts in wealthier countries. However, the correlations also suggest that when groups in poor countries engage in violence, this violence is more likely to spiral into an all-out war, possibly because once the resource costs of starting a rebellion have been overcome, the opportunity costs associated with an all-out war are likely lower in a poor country than in one that is rich. While the robustness of these correlations needs to be subjected to further analysis, the reliance on country level analysis to estimate group level effects is clearly especially problematic with truncated data.

Discussion and conclusion

Ethnic violence has and continues to cause a great deal of pain and suffering; this much is true. The idea, however, that ethnic groups are inherently violent and that ethnic heterogeneity is necessarily problematic for national peace is incorrect. There are many more peaceful ethnic groups in the world than there are violent ones and ethnic heterogeneity is not as clearly a factor in raising the propensity for a minority group to rebel against its state as has been previously estimated. Indeed, ethnic heterogeneity is not dangerous and ethnic war is not as prevalent as has been widely argued. We suggest in this paper that one reason for these misperceptions is the lack of a representative sample of ethnic groups. In particular we highlight problems in inference when truncated group data are erroneously used to generalize about probable group behavior.

In this paper, we acknowledge the difficulty in constructing a list of ethnic groups and build on a frame of commonly recognized groups at a given point in time. Clearly the AMAR sample frame is not comprehensive and the precise boundaries of included groups depend on the research question at hand. Nonetheless, we show that the list of "socially relevant" ethnic groups

in the AMAR frame is consistent with a number of other recent efforts at outlining the universe of ethnic groups that are socially relevant at the national level. Despite substantial differences in methods used in collection, there is great overlap among the datasets. We suggest that this indicates that there is an emerging consensus about the set of socially relevant groups in each country. This is particularly important for cross-national studies that use the ELF measure to describe levels of social heterogeneity. Our new sample data allows us to provisionally conclude that while absolute levels of violence may be higher in heterogeneous countries, there is no reason to believe that heterogeneity is associated with increased group propensity to engage in violence.

For sampling purposes and with respect to answering questions about the causes of political mobilization and violence, the AMAR framework is a substantial improvement over both MAR proper and other more recent collections. For example, we noted earlier that while EPR improves upon the original MAR, it is subject to the same criticism regarding the limitations of the types of groups that are included. Thus, while the EPR framework can be used to test theories about the trajectories of ethnic groups that already are mobilized, like MAR it cannot be reliably used to identify the conditions under which groups become politically relevant or targeted *ab initio.* AMAR does not include any politically relevant criteria for inclusion of an ethnic group in the data. Consequently, some of the ethnic groups in AMAR will be politically relevant and some will not. This is especially important when attempting to sort out the effects of variables related to the selection criteria of either MAR or EPR.

Importantly, however, AMAR is only the enumeration of widely recognized ethnic identities based on an "ethnic social relevance" criteria Looking at underlying ethnic structures – some of which are enumerated in the suggestive lists accompanying the data (see Birnir et al.

2015) – it is clear that many other ethnic configurations exist. It is, therefore, incumbent upon the individual researcher to carefully consider the research question at hand and supplement the data or collect new data on an entirely different configuration of ethnic groups as dictated by the research question at hand. In short, while no dataset on ethnic groups is *the* dataset to answer all pertinent questions, AMAR is a step in the right direction.

Using the weighted AMAR sample, we show that heretofore the frequency of ethnic rebellion against the state is greatly overestimated. In contrast to MAR figures of over two thirds of all groups, we suggest that at most less than a third of ethnic groups ever engage in rebellion against the state. Moreover, because of the fluidity of ethnic boundaries and issues surrounding aggregation, we emphasize that while we have calculated the maximum impact of ethnicity on the probability of ethnic rebellion against the state, the actual likelihood may be even less.

Finally, we compare the bivariate association between group level causes and conflict in the MAR data to the same associations in the weighted AMAR sample. As feared, the MAR sample leaves several group level associations to go undetected or underestimated while rendering others suspect. Equally importantly, we articulate here how the distinct problem of truncation of group data likely causes scholars to erroneously associate aggregate country level characteristics with group behavior. In the weighted AMAR sample – which better accounts for the dependent variable at the group level – most of those aggregate country level associations disappear. It seems likely that greater ethnic heterogeneity in a country is not associated with higher likelihood that any particular group engages in conflict although heterogeneous countries may experience ethnic violence more often. Furthermore, by demonstrating the difference in association when using a dependent measure that captures something akin to a country level total as opposed to one capturing group propensity, we suggest how to reconcile empirically the

conundrum of why developing countries experience greater levels of violence while few groups are able to act upon that grievance. Naturally, this does not suggest absence of ethnic conflict – after all nearly a third of ethnic groups do rebel. What this does demonstrate is that group diversity in and of itself is not a likely significant risk factor for individual ethnic groups. Indeed, with the AMAR correction measure the research community can now, with less worry about selection bias, set about to identify the causes of ethnic violence, a phenomenon less ubiquitous than previously thought, but no less terrifying.

References.

Alesina, Alberto, Arnaud Devleeschauwer William Easterly Sergio Kurlat Romain Wacziarg. 2003. " *Fractionalization*," Journal of Economic Growth, Springer, vol. 8(2): 155-94,

Alonso, Sonia and Rubén Ruiz-Rufino. 2007. "Political representation and ethnic conflict in new democracies." *European Journal of Political Research.* 46: 237–267.

AMAR codebook at http://www.cidcm.umd.edu/mar/amar_project.asp

Birnir, Jóhanna. K., Jonathan Wilkenfeld, James D. Fearon, David D Laitin, Ted R. Gurr, Dawn Brancati, Stephen M. Saideman, Amy Pate, Agatha S. Hultquist. 2015. Socially relevant ethnic groups, ethnic structure, and AMAR. *Journal of Peace Research.* 52 (1):110-115

Birnir, Jóhanna K. 2007. *Ethnicity and Electoral Politics.* Cambridge & New York: Cambridge University Press.

Brancati, Dawn. 2006. "Decentralization: Fueling the fire or dampening the flames of ethnic conflict and secessionism." *International Organization.* 60(3): 651–685.

Brancati, Dawn. 2009. *Peace by Design: Managing Intrastate Conflict through Decentralization*. Oxford: Oxford University Press.

Cederman, Lars-Erik, Brian Min and Andreas Wimmer. 2010. Why Do Ethnic Groups Rebel? New Data and Analysis. *World Politics.* 62(1):87-119

Cetinyan, Rupen. 2002. "Ethnic Bargaining in the Shadow of Third-Party Intervention." *International Organization.* 56(3): 645-677.

Cederman, Lars-Erik, Kristian Skrede Gleditsch and Halvard Buhaug. 2013. *Inequality, Grievances and Civil War.* Cambridge: Cambridge University Press.

Chandra, Kanchan. 2012. *Constructivist Theories of Ethnic Politics*. Oxford: Oxford University Press.

Chandra, Kanchan and Steven Wilkinson. 2008. Measuring the Effect of Ethnicity. *Comparative Political Studies.* 41(4/5): 515-563.

Chromy James R. and Savitri Abeyasekera 2005. "Chapter XIX: Statistical analysis of survey data." In *Studies in Methods Household Sample Surveys in Developing and Transition Countries Section E: Analysis of Survey Data.* Department of Economic and Social Affairs Statistics Division. New York: United Nations. At http://unstats.un.org/unsd/hhsurveys/sectione_new.htm

Collier , Paul and Anke Heoffler. 2004. Greed and Grievance in Civil War. *Oxford Economic Papers.* 56:563-595.

Fearon, James D. and David D. Laitin. 1996. Explaining Interethnic Cooperation. *American Political Science Review*. 90(4): 715-35.

Fearon, James D. 2006. "Ethnic Mobilization and Ethnic Violence." In Donald A. Wittman and Barry R. Weingast eds. *The Oxford Handbook of Political Economy*. Oxford, New York: Oxford University Press.

Fearon, James D. and David D. Laitin. 2003. Ethnicity, Insurgency, and Civil War. *American Political Science Review*. 97(1): 75-90.

Fearon, James D. and David D. Laitin. 2002. "Ethnicity, Insurgency and Civil War." (research project) (http://www.stanford.edu/group/ethnic/).

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth.* 8(2):195-222.

Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics.* Michigan: University of Michigan Press.

Gurr, T. R., and Will H. Moore. 1997. "Ethnopolitical Rebellion: A Cross-sectional Analysis of the 1980s with Risk Assessments for the 1990s," *American Journal of Political Science.* 41(4): 1079-1103.

Hug, Simon. 2013. The Use and Misuse of the "Minorities at Risk" Project. *Annual Review of Political Science*. 16(15):1–18

Hug, Simon. 2010. The Effect of Misclassifications in Probit Models: Monte Carlo Simulations and Applications. Political Analysis. 18(1): 78-102.

Hug, Simon. 2003. Selection bias in comparative research. The case of incomplete datasets. *Political Analysis*. 11(3):255–74.

Kalton, Graham. 1983. "Models in the practice of survey sampling." *International Statistical Review.* 51:175–188.

Kalton, Graham and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics*. 19(2):81-97.

National Election Study at electionstudies.org

Olzak, Susan. 2006. The Global Dynamics of Race and Ethnic Mobilization. Stanford: *Stanford University Press.*

Reagan, Patrick M. and Daniel Norton 2005. Greed, Grievance, and Mobilization in Civil Wars. *Journal of Conflict Resolution*. 49(3):319-336.

Saideman, Stephen M, David J. Lanoue, Michael Campenni, Samuel Stanton. 2002. Democratization, political institutions, and ethnic conflict: a pooled time-series analysis, 1985–1998. *Comparative Political Studies.* 35(1):103–29

Shively, W. Phillips. 2006. "Case Selection: Insights from Rethinking Social Inquiry." *Political Analysis.* 4(3):344-347.

Stoop, Ineke, Jaak Billiet, Achim Koch and Rory Fitzgerald. 2010. *References, in Improving Survey Response: Lessons learned from the European Social Survey*. John Wiley & Sons, Ltd, Chichester, UK.

Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp Hunziker, and Luc Girardin. 2015. Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Data Set Family. *Journal of Conflict Resolution* 59(7): 1327-1342.

Walter, Barbara F. 2006. "Information, Uncertainty, and the Decision to Secede." *International Organization.* 60(1): 105-135.

Weidman, Nils. 2016. A Closer Look at Reporting Bias in Conflict Event Data. *American Journal of Political Science.* 60(1): 206–218

Wimmer Andreas, Lars-Erik Cederman, and Brian Min. 2009. "Ethnic politics and armed conflict. A configurational analysis of a new global dataset." *The American Sociological Review.* 74(2): 316-337.

Öberg, Magnus. 2002. "Minorities Not 'at Risk': A control group for use with Minorities at Risk data." Dataset. Department of Peace and Conflict Research, Uppsala University.

Online Data Appendix

The objective of this data appendix is to elaborate on the technical details of the paper. Below we outline the objective and details of the simulation that was constructed to explore the effects of a particular type of data truncation. Next we discuss treatment of case overlap in the AMAR sample frame. Finally, we outline the construction of the sampling weights. For both the simulation and weights we also make available the code (in R and STATA) so that those interested can manipulate the simulation and weights to explore the effects of alternate specifications.

Simulation

As noted in the paper two types of selection biases are most extensively discussed in the literature. The first is selection on the dependent variable where only (or mostly) cases where the outcome occurs are selected. This is the type of bias discussed extensively by, for example, Geddes (2003) and Shively (2006). Another type of selection issue is recording bias where all cases are recorded but a number of outcomes are recorded incorrectly. Among others, Hug (2010) and Weidman (2016) explore the effects of reporting bias.

The type of bias we are most interested in is a third variant where the data are unintentionally truncated - likely with respect to the dependent variable. Alternatively, the data are intentionally truncated by way of selection criteria but the concern raised in the literature is that the selection criteria may be related to the dependent variable (Hug 2013). This is the type of bias we suspect plagues the original MAR data and perhaps other data on ethnic groups that are selected in similar ways. Specifically, we are interested in finding out what happens when data include most values where the outcome occurs (as there is little evidence of reporting bias on violence in MAR) but exclude many though not all cases where the outcome does not occur.

In particular, we are curious about the effect that this type of bias has on coefficients of relationships between independent and dependent variables.

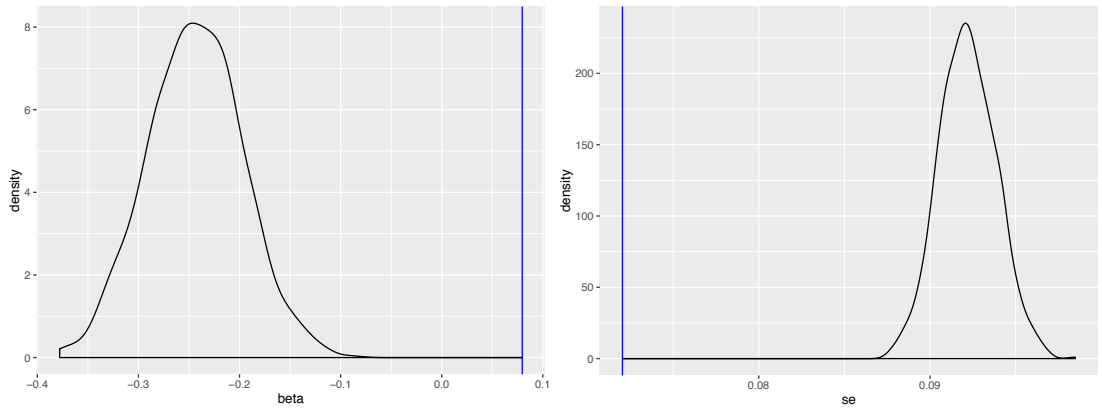
To examine the effects of truncation on estimates of the relationship between the independent and dependent variables, we constructed the following simulation. First, we randomly generated a thousand cases of a dependent variable each half taking on the values of either 0 or 1. We then generated a mostly random independent variable partially correlated with the dependent variable. Specifically, we set 90 percent of the values of the independent variable to draw on a random normal distribution with an average of 0 and a standard deviation of 1. Then, for 10 percent of its values the independent variable was drawn on the dependent variable. This last condition creates a correlation between the independent and the dependent variable. Lastly, we created a selection variable. The selection variable is a sum of the values of the dependent and independent variable and some random noise. Because the independent variable is centered on 0 several of its values are negative. When summed with the dependent variable values of 0s and 1s, selection variable values will more often be negative when the dependent variable is 0. Therefore, we specify that cases be dropped whenever values of the selection variable are below -.5.

We ran our simulations using these variables. Because the objective of each simulation is to compare the effect of truncation to a "true" relationship in the data, in each run we first picked a comparison relationship between given independent and dependent variables generated as per the above description. Next we ran the simulation where we generated 1000 versions of the selection variable used to truncate the data. In other words, in every run the data were truncated on the same principle, described above, but the numbers differed in each run because some were randomly generated. We then compared the distribution of coefficients and standard errors from

the relationship in the truncated data to the relationship between variables in the "true" data. In sum, we found that when compared to results from the "true" un-truncated data, coefficients estimating relationships in the truncated data varied substantially in size, sometimes even changing signs when compared to the "true" correlation. Furthermore, in the truncated data we found standard errors to be invariably larger than in the "true" data though this sometimes rendered correlations more significant and sometimes less, depending on the corresponding size of the coefficient from the truncated data.

Figures A1-A2 plot the results of one of our comparison. In the figures the "true" relationship used for comparison is *not* statistically significant, with a coefficient of 0.080 and a standard error of 0.072 shown with a blue vertical line in each plot respectively. In contrast the relationship between the independent and dependent variable in the simulated data appears as strongly significant with an average coefficient of -0.225 and a standard error of 0.091.

Figures A1 and A2: Coefficients and Standard Errors. Comparison Between a Relationship in A Complete Data (Vertical Line), and Relationships in Truncated Data 1000 Runs (Distribution).



Group overlap and coded groups from the segment that is NEW in the AMAR data.

As noted in the paper, under almost any reasonable operationalization ethnic group markers overlap and cross cut in ways that state designations do not. This problem of overlap has befuddled all those who have sought a complete listing of ethnic groups. Birnir et al. (2015) note that the AMAR project conceptualizes identities as overlapping. However, in its initial iteration (described herein), it "adopts the simplifying assumption of mutual exclusiveness of all groups so that there is minimal overlap in population figures for each group" (2015). For AMAR, the identity category adopted for each country is linked to the dominant way people condition their behavior.[27] In India, for example, the data are organized around religion, caste and tribe while

[27] In the AMAR sample frame some original MAR groups were split, merged or dropped in the AMAR enumeration. These Cases (Country – MAR: AMAR) include: Bolivia -- Highland Indigenous: Aymara and Quechua; Zanzibar – Zanzibaris: Zanzibar Africans/Shirazi and Zanzibar Arabs; India – Scheduled Tribes Of India: Scheduled Tribes of East, North, Northeast,

regional identities such as Bengali and Marathi are not coded. Meanwhile for Nigeria, the data are organized around tribe with religious identities (Muslim, Traditional, and Christian) omitted. In choosing which set of categories to list, Birnir et al. (2015) relied on country sources to fill out an initial list, to be modified by future scholars, of socially relevant identity groups capturing most if not all of a country's population. Importantly, even within a country the socially relevant identity may vary between groups. For example, in China Islam is likely a relevant identity for some minorities whereas religion may not be a relevant identity choice for the majority Han Chinese.

As scholars work to fill in the complete identity profile for each country, they will have to take greater notice of the relevant list of identities for each group and overlap within a group. The implication of overlap for the sample frame and the sampling calculations in this paper is that in order to have the sum of all groups in a country be around 100% of the population (which would be violated, for example, if we counted all language and all caste groups separately in India), in many cases we needed to choose an identity axis that is predominant in ethnic relations

---

South, and West India; Switzerland – Foreign Workers: Spanish, Portuguese and former Yugoslavs. Mozambique – Makonde/Yao: Makonde , Yao, and Makua. Switzerland – Jurassians and French Speakers: French Speakers. West, East and Germany Turks: Turks. Yugoslavia – Kosovo Albanians: Serbia - Albanians. Socialist Federal Republic of Yugoslavia - Croats: dropped because don't meet population threshold criteria. South Korea - Honamese: dropped because don't meet population threshold criteria. South Vietnam – Chinese: dropped because don't meet population threshold criteria. South Vietnam – Montagnards: dropped because don't meet population threshold criteria. Also, the Turkmen in China are now referred to as the Uygurs in AMAR.

in a country. However, we can make all other axes sub-groups of the dominant axis (and thus for India, each caste and religious group will have language sub-groups), so that scholars can easily use our dataset to alter the identity axis to be analyzed. This of course is only a partial solution. Although tractable in the long term, the issue of overlap presents challenges to students of ethnic groups.

Weights

We note in the paper that weighting is common in survey analysis, where sample segments often over- or under-represent particular population segments (men and women for example). Following common convention the weights we use are the inverse of an individual observation's probability of selection into the sample from the population $W_i = \frac{1}{P_i}$. Thus, sample weights are inflation or deflation factors that allow a sample unit to represent the number of units in the survey population that are accounted for by the sample unit to which the weight is assigned.

Our data are separated into two segments - 288 MAR groups and 74 NEW groups from the AMAR sample frame. Together these segments (365 groups combined) make up the overall AMAR sample. Because individual observations in each segment have unequal probabilities of being selected into the sample, our weights account separately for the probability of individual selection in each segment. In other words $W_m = \frac{N_m}{n}$ where m is either segment (MAR or NEW), n_m denotes the share of the population in each segment as per the AMAR sample frame where $n = \sum_{m=1}^{i} n_m$ = total sample size, $N_m =$ population size of segment m, m = 1,2,3,….i, and $N = \sum_{m=1}^{i} N_m$ =total population size. Because included MAR groups represent the full list of current MAR in the AMAR sample frame (population), each MAR group represents one group

from the population $W_{m(MAR=1)} = \frac{288}{288}$[28]. The inverse sampling probability for MAR cases is, therefore, 1. The 74 NEW groups, in turn, constitute a random sample from the remaining 911 AMAR groups. Each of these groups is also weighed by its inverse sampling probability $W_{m(NEW=1)} = \frac{911}{74}$. Our weighting assignment for groups in NEW also takes into account region and population strata so as not to replicate known problems from the MAR segment of the sample.[29]

One criticism of this type of commonly used weights is that the potentially large distribution of resultant weighting adjustments can inflate the variance in sample responses (Kalton and Flores 2003). Responding to this criticism we conservatively scale our weights for the sample size with respect to the Universe (365/1202)[30] multiplying each weight with this number, thereby reducing the variance in the weights significantly.[31]

[28] Aside from outbidding (Horowitz, 1985) intragroup fighting among dominant majorities has not been a focus of ethnic conflict studies, thought this is changing (Hultquist n.d.) Consequently we exclude continually politically dominant groups both from calculation of weights and from the analysis. Continually politically dominant refers to a group that consistently controls or is a senior partner in the executive in democratic countries or the equivalent in authoritarian countries. We used EPR coding for this information supplementing with country specific accounts where EPR did not code these data.

[29] The weights are assigned to a cross-section of the data and do not account for features pertaining only to panel data. This is a simplification that bears further scrutiny especially with respect to uneven missing-ness of data between years.

[30] Note that the actual numbers in the data differ slightly after we account for dominant groups.

[31] We thank Rick Valliant for this and other helpful suggestions about weighting.

Weights are most commonly used in descriptive analysis but recent literature notes that weights can be used in inferential analysis also (Chromy and Abeyasekera 2005). [32] Throughout the duration of this project we have experimented with a variety of different weights including straight cell weights, cell weights with reduced variance as described above, and raking weights. By and large in comparison to analysis done on uncorrected samples (such as the original MAR) our observation is that sample correction in conjunction with inclusion of weights is helpful. It is less clear how the benefits and drawbacks of each weighting scheme compare but this remains a topic for further study.

[32] Strijbis (2013) suggests an alternative weighting strategy for fuzzy categories such as ethnic groups where prototypical cases are weighted more heavily than are boundary cases. To use this weighting strategy the proto-typicality of each AMAR case would have to be operationalized. While outside the scope of this study this presents an interesting direction for future research.

Appendix References

Chromy James R. and Savitri Abeyasekera 2005. "Chapter XIX: Statistical analysis of survey data." In *Studies in Methods Household Sample Surveys in Developing and Transition Countries Section E: Analysis of Survey Data.* Department of Economic and Social Affairs Statistics Division. New York: United Nations. At http://unstats.un.org/unsd/hhsurveys/sectione_new.htm

Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics.* Michigan: University of Michigan Press.

Horowitz, Donald. 1985. *Ethnic Groups in Conflict.* Berkeley: University of California Press.

Hug, Simon. 2013. The Use and Misuse of the "Minorities at Risk" Project. *Annual Review of Political Science*. 16(15):1–18

Hug, Simon. 2010. The Effect of Misclassifications in Probit Models: Monte Carlo Simulations and Applications. Political Analysis. 18(1): 78-102.
Explaining Interethnic Cooperation. *American Political Science Review.* 90(4): 715-35.

Hug, Simon. 2003. Selection bias in comparative research. The case of incomplete datasets. *Political Analysis.* 11(3):255–74.

Hultquist, Agatha. N.d. Dissertation: "Bringing ethnicity back in: examining the effects of ethnic majority and minority competition on inter- and intra-ethnic conflict and cooperation." University of Maryland. Typescript.

Kalton, Graham. 1983. "Models in the practice of survey sampling." *International Statistical Review.* 51:175–188.

Kalton, Graham and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics*. 19(2):81-97.

Shively, W. Phillips. 2006. "Case Selection: Insights from Rethinking Social Inquiry." *Political Analysis.* 4(3):344-347.

Strijbis, Oliver. 2013. Prototypical Weighting: Toward a Solution for Macrosociological Comparisons of Fuzzy Cases. Sociological Methods & Research. 42(4): 458-482.

Weidman, Nils. 2016. A Closer Look at Reporting Bias in Conflict Event Data. *American Journal of Political Science.* 60(1): 206–218